Enhancing of DBSCAN based on Sampling and Density-based Separation

Safaa O. Al-mamory

College of Business Informatics

University of Information Technology and Communications

Baghdad, Iraq

salmamory@uoitc.edu.iq

Israa Saleh Kamil

Babylon, Iraq

esraa@itnet. uobabylon.edu.iq

***Abstract*— DBSCAN (Density-Based Clustering of Applications with Noise)is one of the attractive algorithms among density-based clustering algorithms. It characterized by its ability to detect clusters of various sizes and shapes with the presence of noise, but its performance degrades when data have different densities .In this paper, we proposed a new technique to separate data based on its density with a new sampling technique , the purpose of these new techniques is for getting data with homogenous density .The experimental results on synthetic data and real world data show that the new technique enhanced the clustering of DBSCAN to large extent.**

***Index Terms*—Density-based, DBSCAN, different densities, Sampling.**

I. INTRODUCTION

Clustering is the process of collecting set of objects into different classes where the objects in each class have high similarity to each other and low similarity with objects in other classes[1] . Density-based clustering is one of the most important classes in cluster analysis, the main idea in its work is that it forms clusters by collecting objects that are densely connected and separated by sparse regions[2].It contains algorithms like OPTICS(Ordering Points To Identify the Clustering Structure)[3], DENCLUE(DENsity-based CLUstEring) [4], and DBCLASD(Distribution Based Clustering of Large Spatial Databases) [5].

DBSCAN (Density-Based Clustering of Applications with Noise) [6] is an efficient density-based clustering algorithm that detects clusters of different shapes and sizes ,but the using of global density threshold constitute a defect in the clustering result when data contain different densities .Many researchers propose different solutions to solve the problem of DBSCAN with different densities. In this paper, we present a new solution to solve that problem by using a new technique to separate data depending on the density contained in them , then using a new sampling technique to get data with homogenous density , the data resulted from the separation and sampling will be clustered by DBSCAN , and finally , KNN will be applied on the core points resulted from the previews step with the dense data remaining after sampling.

The remaining of the paper is ordered as follows , Section 2 presents related works , a general overview on DBSCAN will be presented in Section 3 ,our proposed algorithm will completely implemented and illustrated in Section 4 , experimental results on synthetic data and real world data were applied and tested in Section 5 , and Section 6 will discuss the main conclusions.

II.RELATED WORKS

Because of DBSCAN considered as one of the most important density-based clustering algorithms, many of

approaches were proposed to enhance its performance to deal with data with different densities.

The concept of shared nearest neighbors of each data point were developed by [7] ,their idea was to compute the number of neighbors that shared by each pair of points .This novel definition of similarity helps in removing noise or outliers ,recognizing core points ,and also creating clusters surrounding the core points .

When using spatial index together with grid technique, the result is GMDBSCAN [8] (Multi-Density DBSCAN Cluster Based on Grid) ,this method uses space dividing technique and consider each grid as a separate part ,then it estimates independent *Minpts* for every grid (part) based on its density ,after that it applies multiple DBSCAN on each grid ,and finally, it uses distance –based method to improve boundaries. By using the density distribution, *Huang et al.* [9] proposed a new algorithm in which the parameter *Eps* with two constraints is calculated for each density distribution , where it gets referenced *Eps* from the must link set, then from these referenced *Eps* it will select representative *Eps* by cannot link constraint and finally, it uses multi-stage DBSCAN to cluster each partition with the representative *Eps*.

Hua et al.[10] presented a new technique where firstly, the space of data is divided into a number of grids.Secondly , the space of data is divided again to get smaller partitions , this division of the space is done according to the one –dimensional or two-dimensional characteristics of the density distribution of the grid and finally , for each partition it applies an improved DBSCAN with different parameters to cluster these partitions respectively.

In the same context, *Ren et al.* [11] proposed a new method called DBCAMM (density based clustering algorithm with Mahalanobis metric) that developed DBSCAN. Firstly , by replacing Euclidian distance by Mahalanobies distance metric ,this metric is associated with the distribution of the data and secondly, by introducing a method to combine sub-clusters by using the information of the density of the sub-cluster.

GRPDBSCAN (Grid-based DBSCAN algorithm with referential parameters) is another solution to the problem of different densities ,where the merging of multi-density based clustering and grid partition technique and the automatic generation of ***Eps*** and ***Minpts*** is performed to enhance DBSCAN [12]**.**

Zhang, Xu and Si [13] introduced a new technique depends on four concepts: Contribution ,grid technique , migration-coefficient, and tree index structure to optimize the performance of DBSCAN to be able to discover clusters with different densities . This optimization is carried out by firstly, using grid technique to reduce the time where the algorithm will be efficient for large databases. Secondly, the optimization of the clustering results is fulfilled by expressing the density of the grid based on the concept of contribution . Thirdly, the improving of the clustering quality will be done by focusing on boundary points using migration coefficient. In M-DBSCAN(Multi density-DBSCA***)*** [14], neighbors didn't found with a constant radius ε,instead the determination of neighboring radius is performed based on the data distribution around the core using standard deviation and mean values. To get the clustering results, M-DBSCAN is applied on a set of core-mini clusters where each core-mini cluster represents a virtual point lies in the center of that cluster. In M-DBSCAN local density cluster is used as an alternative to ε value of DBSCAN , by adding core-mini clusters that have similar mean values with a little difference determined by the standard deviation of the core, the clusters is extended in this algorithm.

III.PRILIMINARY

DBSCAN was firstly proposed by Ester et al. [6] , its work depends on determining two parameters , the first one is the density threshold (*Minpts*) and the other is *Eps* which is the radius within it the point should contain number of neighbors greater than or equal to Minpts. It starts with an

arbitrary point and retrieve all points that are density – reachable from that point within the specified Eps and Minpts [6].It defines a cluster as a maximum set of density-connected data points ,where a point is considered to be core point if it has number of neighbors greater than or equal to Minpts within Eps . All points within the same cluster could be reached by using density-connected concept. DBSCAN is so sensitive to the selection of Minpts and Eps values , where a slightly different setting of them may lead to very different clustering of dataset
[1].

Given a set of objects ,*D*, we say that an object *P* is directly-density reachable from object *q* if *p* is within the ε – neighborhood of *q* , and *q* is a core point .An object *p* is density-connected to object *q* with respect to ε and Minpts in a set of objects *D*, if there is an object $o \in D$ such that both *p* and *q* are density-reachable from *o* with respect to ε and Minpts , these two concepts are illustrated in Fig.1.(a) and Fig.1.(b).

IV. THE PROPOSED ALGORITHM

The proposed algorithm consists of four main steps that are illustrated in Fig 2

.*A. Density-Based Separation*

Many approaches have been used to construct the density levels of the data like using density estimators such as histograms , naïve estimators ,kernel estimators , k-nearest neighbor estimators , and etc . [15].In this paper we propose a new technique that could be used to separate data based on the contrast in density that is contained in data. Here two constraints are applied on data to separate it into two regions : Dense Data (DD) and Sparse Data (SD) . Given a data set D , then the first constraint is the global mean that represents the mean value of the distances for all the points with respect to *kth* nearest neighbors of them and we will refer to it by GM in the remaining Sections of this paper and is computed by Equation.1. The second constraint is the local mean which is the mean value of the distances of one point with respect to *kth* of its nearest neighbors and we will refer to it by LM and this is computed by Equation.2 .

Fig .1: Density concepts (a) density- reachability (b) density-connectivity.

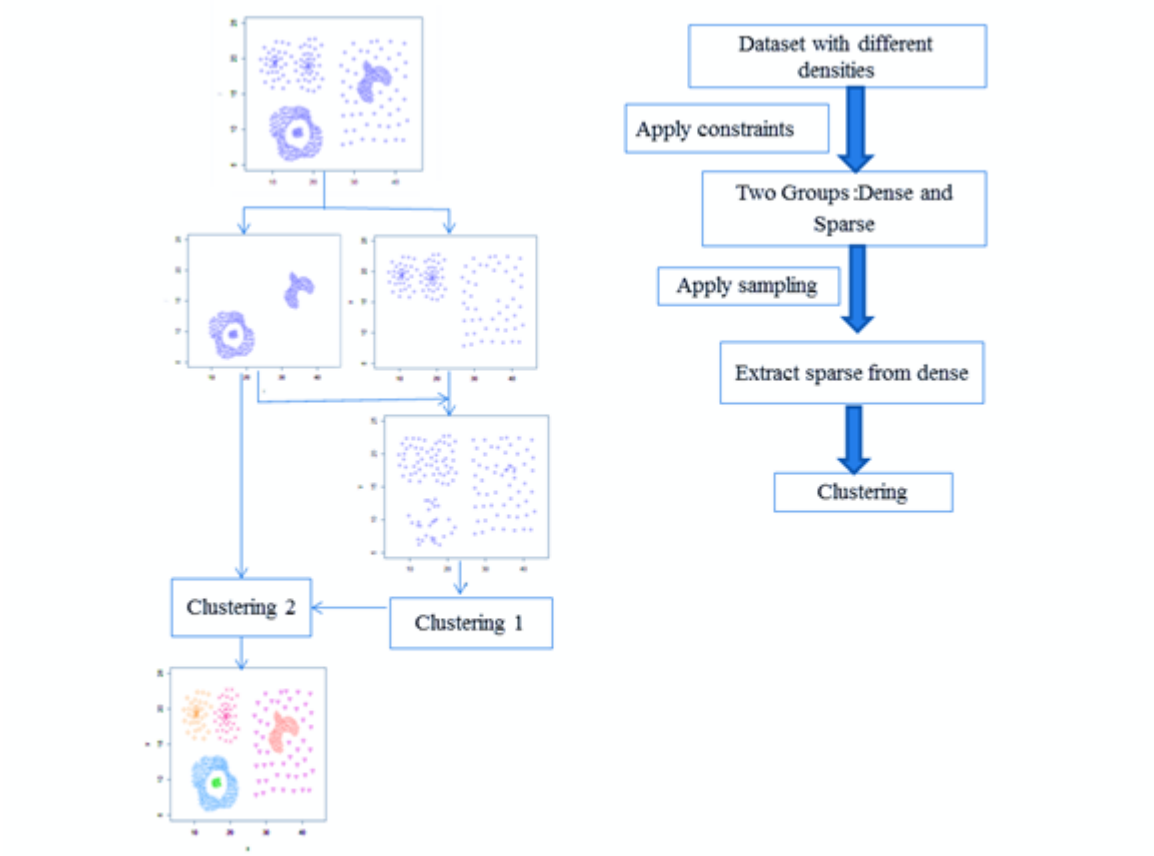

Fig.2: Diagram of the proposed system

$$\mathbf{GM} = \frac{1}{n \times m} \sum_{i=1}^{n} \sum_{j=1}^{m} \boldsymbol{dist}(\boldsymbol{p_i}, \boldsymbol{p_j}) \quad \textbf{(1)}$$

$$\mathbf{LM}(\boldsymbol{p_i}) = \frac{1}{k} \sum_{l=1}^{n} \boldsymbol{dist}(\boldsymbol{p_i}, \boldsymbol{p_l}) \quad \textbf{(2)}$$

Where $dist \neq 0$, m and n represent dimensions of sparse matrix , and k represents number of nearest neighbors .After computing these constraints the data will be separated according to the following definitions :

Definition 1: (Dense point) any point *p* is considered as dense point if it satisfies the following condition

LM(p) < GM(D)

Defenition 2: (Sparse point) any point *p* is considered as sparse point if it satisfies the following condition

LM(P) > GM(D)

B.Density Biased Sampling

A lot of attention was paid to density biased sampling techniques in the last few years .Density biased sampling contains a special case of uniform sampling , some of the proposed density biased sampling algorithms are used for important applications like clustering where dense regions are oversampled and light regions are under-

sampled because uniform sampling misses the small clusters [16].

In this paper , we propose a new density biased sampling technique that is used to get sparse data from dense data . the purpose of this step is to provide DBSCAN with data that have homogenous density distribution where DBSCAN works well . The proposed method does not need to determine the size of the sample because it will be determined automatically depending on the density of the population , Algorithm 1. Describes the procedure of the new sampling technique.

Algorithm 1: Density Biased Sampling (DD,SD)

Input: the set of dense points DD and the set of sparse points SD.

Output: subset of dense points .

1.Begin

2. Compute GM(DD).

3. Compute GM(SD).

4. While (GM(DD) < GM(SD))

5. select point ***p*** from DD randomly .

6. remove ***p*** from DD.

7. re-compute distance matrix and sparse matrix for DD.

8. compute new GM(DD).

9. END While .

10.END.

C.Clustering of SD with DBSCAN

In this step DBSCAN will be applied on data constructed from the previews step (Subsection 4.2) where data become with homogenous density distribution . DBSCAN gives good clustering when it is applied on data free of different densities , so here the final number of clusters will be determined by DBSCAN clustering . Because of that the determination of the parameters of DBSCAN is not an easy process , then k-dist plot is used here to determine these parameters by computing k-nearest neighbors for each point and sort them by ascending and then plotting them [17], As it is well known that DBSCAN produce core points and noise points , both of these points will be passed to the next step to be clustered with the dense points that are remained from sampling process by KNN.

D.Clustering of DD and Core Points by KNN

The final clustering process will be performed by applying KNN on the dense data that are remained from sampling with core points resulted from DBSCAN , noise points will also clustered in this step , so the final clustering will not produce noise points absolutely . This clustering is carried out by computing the distance between a given dense point with all the core points and consequently , the dense point will assign to the cluster that contains the closest core point to it, for noise points the same process will be applied . Finally, all the points will be assigned

to a particular cluster to get the whole number of clusters of data.

E.Evaluation Metrics

One of the most common evaluation measures is F-measure that is used for the evaluation of classification and clustering tasks.

1.F-Measure

The number of correct results divided by the number of all returned results is called Precision and the number of correct results divided by the number of results that should have been returned is called recall [14], these two definitions are described by Equation 3. and Equation 4.

$$\boldsymbol{precision}\,(\boldsymbol{C},\boldsymbol{G}) = \sqrt{\frac{\Sigma_{ij}\binom{n_{ij}}{2}}{\Sigma_{ij}\binom{n_{ij}}{2}+\Sigma_{j}\binom{b_j}{2}+\Sigma_{ij}\binom{n_{ij}}{2}}} \quad \textbf{(3)}$$

$$\boldsymbol{Recall}(\boldsymbol{C},\boldsymbol{G}) = \sqrt{\frac{\Sigma_{ij}\binom{n_{ij}}{2}}{\Sigma_{ij}\binom{n_{ij}}{2}+\Sigma_{i}\binom{a_i}{2}-\Sigma_{ij}\binom{n_{ij}}{2}}} \quad \textbf{(4)}$$

Where n_{ij} represents the number of points of class i in cluster j, b_j represents number of points in cluster j , and a_i represents number of points in class i .To compute the score, F-Measure or F-score[17] used both of precision and recall. F-Measure represents a harmonic mean for both of precision and recall where the value closer to 1 refers to good clustering and the value closer to 0 indicates the opposite and that measure could be computed by Equation.5.

$$\boldsymbol{F-Measure}(\boldsymbol{C},\boldsymbol{G}) = \frac{2*\boldsymbol{precision}(\boldsymbol{C},\boldsymbol{G})*\boldsymbol{Recall}(\boldsymbol{C},\boldsymbol{G})}{\boldsymbol{precision}\,(\boldsymbol{C},\boldsymbol{G})+\boldsymbol{Recall}(\boldsymbol{C},\boldsymbol{G})} \quad \textbf{(5)}$$

V. EXPERIMENTAL RESULTS

For the test purposes, two-dimensional synthetic dataset known as Compound dataset from[18] is used to confirm the benefits of the new proposed algorithm, where the synthetic dataset have the characteristics of difference in shape, size, and density which is the focus of this paper. Fig .3.(a) shows the k-dist plot for all points of the Compound data and Fig .3.(b) shows the k-dist of sparse data points only that are constructed from Compound dataset , this plot is used to determine the values of the parameters of DBSCAN which is used to cluster the sparse data.

Figure 4. (a) illustrate the clustering results of DBSCAN on Compound dataset and there is clear inferior performance of DBSCAN with this type of, in Fig.4(b) there is enhancing for the clustering results of DBSCAN with the proposed algorithm .

a **b**

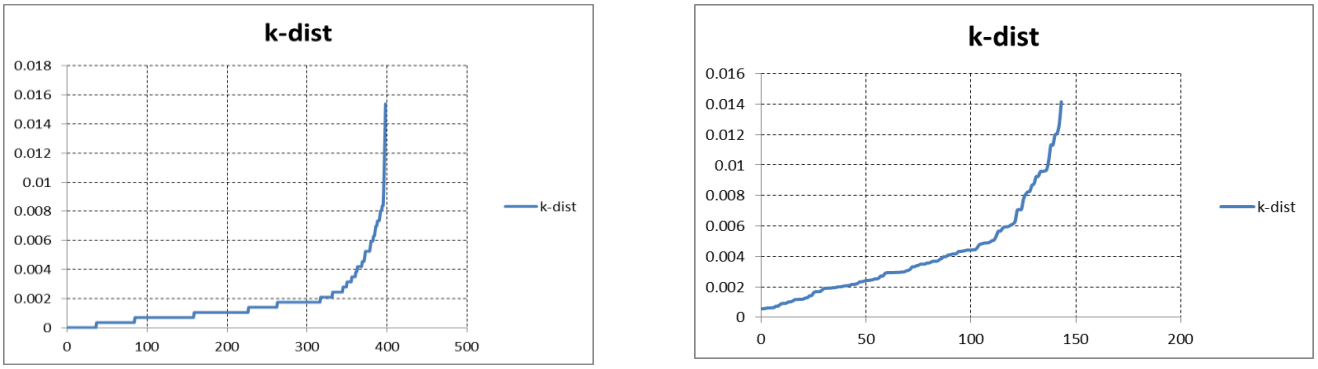

Fig.3: K-dist plot for Compound dataset (a) original data (b) sparse data

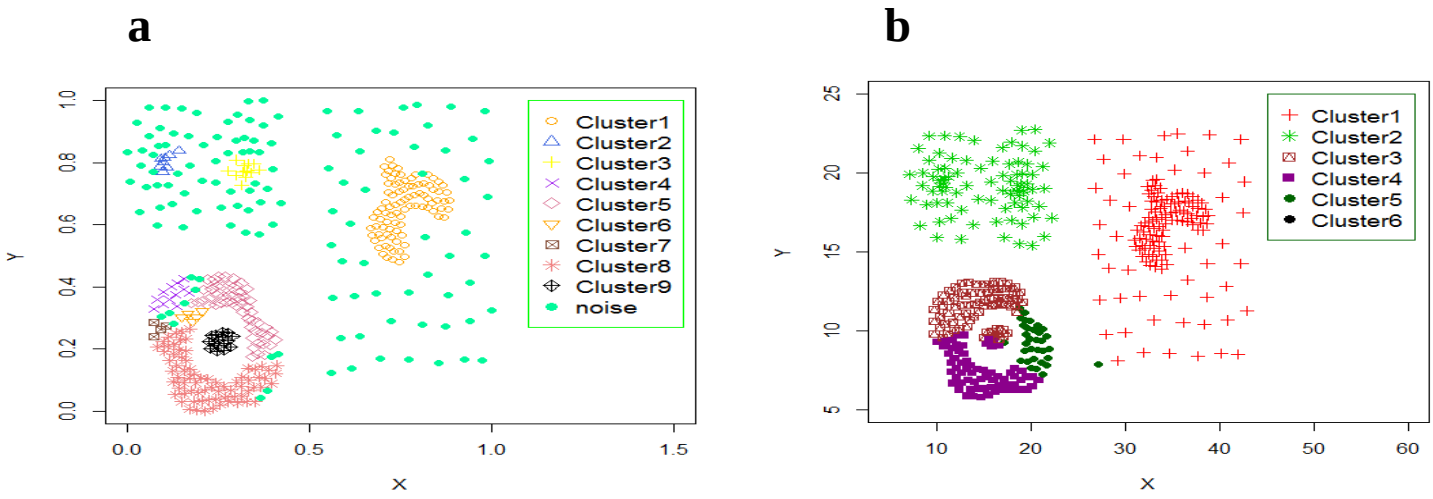

Fig.4: Clustering results of (a) DBSCAN (b) Proposed algorithm

The proposed algorithm also was tested on real world multidimensional dataset represented by Blood Transfusion service center from UCI, this dataset have the property of difference in density . By using F-measure , we get a comparison between DBSCAN and the proposed algorithm on the synthetic and real world datasets that are mentioned previously . The comparison shows that the values of precision , recall , and F-measure for the proposed algorithm are better than that of DBSCAN for both of synthetic and real world datasets as that illustrated in TABLE 5. The value of F-measure is computed based on the confusion matrix that is illustrated in TABLE.1,TABLE.2,TABLE.3,and TABLE.4 for clustering of Compound dataset with DBSCAN, clustering of Compound dataset with proposed algorithm, clustering of Blood Transfusion service center with DBSCAN, and clustering of Blood Transfusion service center with proposed algorithm respectively.

TABLE 1: Confusion matrix of clustering with DBSCAN for Compound dataset.

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Total
Cluster 1	0	91	0	0	0	0	91
Cluster 2	0	0	10	0	0	0	10
Cluster 3	0	0	0	12	0	0	12
Cluster 4	0	0	0	0	12	0	12
Cluster 5	0	0	0	0	53	0	53
Cluster 6	0	0	0	0	5	0	5
Cluster 7	0	0	0	0	4	0	4
Cluster 8	0	0	0	0	73	0	73
Cluster 9	0	0	0	0	0	16	16
Total	0	91	10	12	147	16	276

TABLE 2: Confusion matrix of clustering with proposed algorithm for Compound dataset.

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Total
Cluster 1	49	92	0	0	0	0	141
Cluster 2	0	0	38	45	0	0	83
Cluster 3	0	0	0	0	66	0	66
Cluster 4	0	0	0	0	61	16	77
Cluster 5	1	0	0	0	31	0	32
Total	50	92	38	45	158	16	399

TABLE 3:Confusion matrix of the proposed algorithm by for Blood Transfusion service center dataset.

	Class 1	Class 2	Total
Cluster 1	173	570	743
Cluster 2	5	2	7
Total	178	572	750

TABLE 4 :Confusion matrix of DBSCAN for Blood Transfusion service center dataset.

	Class 1	Class 2	Total
Cluster 1	152	470	622
Cluster 2	4	3	7
Cluster 3	3	9	12
Cluster 4	0	4	4
Cluster 5	1	11	12
Cluster 6	0	4	4
Cluster 7	1	4	5
Cluster 8	1	16	17
Cluster 9	0	4	4
Cluster 10	0	4	4
Total	162	529	691

TABLE 5:Comparison between DBSCAN and the proposed algorithm .

Dataset	DBSCAN			Proposed Algorithm		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Compound	0.5773	0.5588	0.5678	0.5254	0.7692	**0.6243**
Blood Transfusion service center	0.5280	0.7985	0.6356	0.5302	0.9887	**0.6902**

VI. CONCLUSIONS

In this paper ,a new technique was presented , in which the data is separated into groups based on the density that determined by applying some constraints .In the clustering process ,the data was separated perfectly , then new sampling technique was applied in order to reduce the density of dense data and obtain data with only one density distribution (sparse data), the results of sampling were very effective .The experiments performed on synthetic data show that the proposed algorithm enhanced the performance of DBSCAN on data with different densities as the F-measure proves that .

REFERENCES

[1] J. Han, M. Kamber, and J. Pei, " Data Mining:Concepts and Techniques",2nd Edition , *Morgan Kaufmann Series in Data Management Systems* . Elsevier, 2006.

[2] W. Loh and Y. Park, "A Survey on Density-Based Clustering Algorithms," *springer*, pp. 775–780, 2014.

[3] M. Ankerst, M. M. Breunig, H. Kriegel, and J. Sander, "OPTICS : Ordering Points To Identify the Clustering Structure," *SIGMOD '99 Proc. 1999 ACM SIGMOD Int. Conf. Manag. data*, vol. 28, no. 2, pp. 49–60, 1999.

[4] A. Hinneburg and D. A. Keim, "An E cient Approach to Clustering in Large Multimedia Databases with Noise," *American Association for Artificial Intelligence*, 1998.

[5] X. Xu, M. Ester, H. Kriegel, and J. Sander, "A Distribution-Based Clustering Algorithm for Mining in Large Spatial Databases ", Proceedings of 14th International Conference on Data Engineering (ICDE ' 98) D-80538 München 3 , pp. 324–331, 1998.

[6] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," *Second International Conference on Knowledge Discovery and Data Mining*, pp. 226–231, 1996.

[7] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, " *Cluster Analysis "*, 5th Edition ,*John Wiley & Sons* ,vol. 14, pp.348, 2011.

[8] C. Xiaoyun, M. Yufang, Z. Yan, and W. Ping, "GMDBSCAN: Multi-density DBSCAN cluster based on grid," *IEEE International Conference on e-Business Engineering*, pp. 780–783, 2008.

[9] T. Q. Huang, Y. Q. Yu, K. Li, and W. F. Zeng, "Reckon the parameter of DBSCAN for multi-density data sets with constraints," *2009 International Conference on Artificial Intelligence and Computational Intelligence, AICI 2009*, vol. 4, no. 1996, pp. 375–379, 2009.

[10] Z. Hua and W. Zhenxing, "Clustering algorithm based on characteristics of density distribution,"

2010 2nd International Conference on Advanced Computer Control (ICACC), vol. 2, pp. 431–435, 2010.

[11] Y. Ren, X. Liu, and W. Liu, "DBCAMM: A novel density based clustering algorithm via using the Mahalanobis metric," *Applied Soft Computing Journal*, vol. 12, no. 5, pp. 1542–1554, 2012.

[12] H. Darong and W. Peng, "Grid-based DBSCAN Algorithm with Referential Parameters," *International conference on Applied Physics and Industrial Engineering , Physics . Procedia*, vol. 24, pp. 1166–1170, 2012.

[13] L. Zhang, Z. Xu, and F. Si, "GCMDDBSCAN: Multi-density DBSCAN Based on Grid and Contribution," *2013 IEEE 11th International Conference on Dependable, Autonomic and Secure Computing.*, pp. 502–507, 2013.

[14] A. Amini, H. Saboohi, T. Herawan, and T. Y. Wah, "MuDi-Stream: A multi density clustering algorithm for evolving data stream," *Journal of Network and Computer Applications*, 2014, pp. 1–16.

[15] C. R. Palmer and C. Faloutsos, "Density biased sampling," *Proceedings of the 2000 ACM SIGMOD international conference on Management of data - SIGMOD '00*, pp. 82–92, 2000.

[16] L. Duan, D. Xiong, J. Lee, and F. Guo, "A Local Density Based Spatial Clustering Algorithm with Noise," *2006 IEEE International Conference on Systems, Man and Cybernetics*, no. October, pp. 4061–4066, 2006.

[17] C. J. V. R. B. Sc, " INFORMATION RETRIEVAL" , 2nd Edition. Newton, MA, USA: Butterworth-Heinemann , 1979.

[18] C.T., Zahn, "Graph-theoretical methods for detecting and describing gestalt clusters." *IEEE Transactions on Computers*, 1971, 100(1): pp. 68-86.